Análise da autocorrelação espacial de dados do Twitter

Spatial autocorrelation analysis of Twitter data

Fábio Bays de Araujo[1], Sidgley Camargo de Andrade[2]

RESUMO

Este estudo conduz uma análise da autocorrelação espacial de *tweets* públicos e georreferenciados no município de São Paulo, Brasil, visando verificar a dependência espacial do número absoluto de *tweets* e da densidade de *tweets* ao longo dos distritos. Para a análise, foi usado um conjunto de dados disponível em repositório aberto. Resultados mostram que há autocorrelação espacial global positiva forte para o número absoluto de *tweets* e fraca para a densidade de *tweets*. Com relação à autocorrelação local, os resultados foram similares entre o número absoluto de *tweets* e a densidade de *tweets*; com autocorrelações espaciais localizadas tanto positivas quanto negativas.
PALAVRAS-CHAVE: autocorrelação espacial; ciência de dados geográfica; tweets.

ABSTRACT

This study conducts an analysis of the spatial autocorrelation of public geotagged tweets that fell within São Paulo city, Brazil. The analysis aims to verify the dependence between the values of absolute number of tweets and of density of tweets across the districts. A dataset available in an open repository was used in the analysis. Results show strong positive global spatial autocorrelation for the absolute number of tweets and weak for the density of tweets. For the local autocorrelation, the results are similar between the absolute number of tweets and density of tweets; with both positive and negative localized spatial autocorrelation.
KEYWORDS: spatial autocorrelation; geographic data science; tweets.

INTRODUÇÃO

Com o advento e democratização das redes sociais e da Internet, um grande número de conteúdo gerado por usuário (UGC - *User-Generated Content*) tem sido produzido. Esse tipo de dado tem se tornado precioso para os pesquisadores como uma fonte de dados complementar às fontes de dados tradicionais (por exemplo, sensores e *surveys*). Uma das fontes de UGC são as plataformas de redes sociais como o X/Twitter, cujos dados podem conter uma localização geográfica (longitude e latitude). Dados de redes sociais têm sido utilizados em estudos de análise da percepção humana sobre fenômenos do mundo real (Yang W.; Mu; Shen, 2015), da obesidade e concentração de restaurantes de fast-food (Chen; Yang X., 2014), da localização de infraestrutura verde apreciada pelos habitantes de um município (Guerrero *et al.*, 2016) e da taxa de visitação de parques e atrações recreativas (Wood *et al.*, 2013).

Uma das análises fundamentais em dados contendo informações geográficas é a da autocorrelação espacial. A autocorrelação espacial diz respeito ao quanto o valor de uma variável (por exemplo, uma característica em uma região) é influenciada pelo valor das regiões vizinhas para a mesma variável (Rey; Arribas-Bel; Wolf, 2020). Em outras

[1] Bolsista da Universidade Tecnológica Federal do Paraná. Universidade Tecnológica Federal do Paraná, Toledo, Paraná, Brasil. E-mail: fabiobays@alunos.utfpr.edu.br. ID Lattes: 0071612259604427.
[2] Docente no Curso de Engenharia da Computação/COENC. Universidade Tecnológica Federal do Paraná, Toledo, Paraná, Brasil. E-mail: sidgleyandrade@utfpr.edu.br. ID Lattes: 2230323637134843.

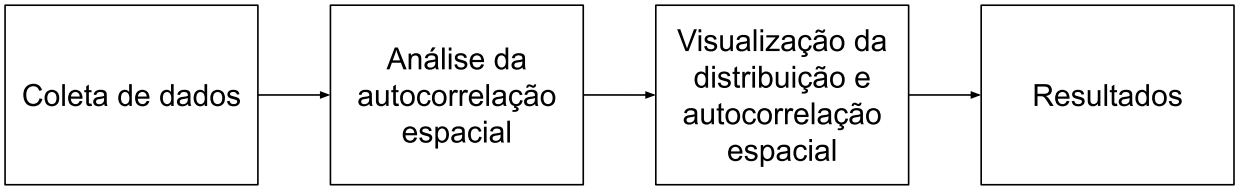
palavras, a autocorrelação espacial serve para indicar se há dependência entre os valores para uma variável mapeada em um espaço.

O presente estudo conduz a análise da autocorrelação espacial de *tweets* públicos localizados no município de São Paulo. Duas variáveis são comparadas, o número absoluto de *tweets* e a densidade de *tweets* (número de *tweets* por habitante).

METODOLOGIA

A Figura 1 apresenta a metodologia aplicada na análise da autocorrelação espacial dos *tweets*.

Figura 1 - Metodologia da análise da autocorrelação espacial.



Fonte: elaborado pelo autor (2023).

COLETA DE DADOS

Foram utilizados dados históricos coletados entre 2016 e 2017 via Twitter *Streaming* API e obtidos do estudo de Andrade *et al*. (2021). O conjunto de dados possui 1.214.611 *tweets* públicos e georreferenciados com o parâmetro de localização do município de São Paulo, Brasil. Os dados estão disponíveis em um repositório público disponível em https://doi.org/10.6084/m9.figshare.12921974.

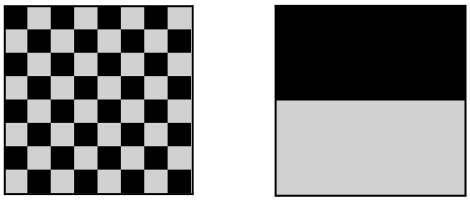
ANÁLISE DA AUTOCORRELAÇÃO ESPACIAL

Duas importantes estatísticas espaciais são a de autocorrelação espacial global e local, que mostram se há uma tendência de agrupamento de regiões similares ou distintas entre si, de maneira geral ou localizada, dado uma característica (Rey; Arribas-Bel; Wolf, 2020).

Embora não sejam as únicas estatísticas de autocorrelação, os índices de Moran global e local são índices amplamente utilizados em estudos que procuram avaliar a correlação de variáveis dentro do espaço urbano.

Para a aplicação dos índices, calcula-se a autocorrelação de uma característica entre regiões vizinhas. O índice de Moran global varia no intervalo [-1,+1], enquanto o índice de Moran local varia no intervalo ($-\infty$ a $+\infty$). Por um lado, quanto menores os índices, maior a autocorrelação espacial negativa. Isso denota um maior agrupamento entre regiões com características distintas. Por outro lado, quanto maiores os índices, maior a autocorrelação espacial positiva. Nesse caso, regiões com características similares estão mais agrupadas. A Figura 2 exemplifica ambos casos.

Figura 2: Autocorrelação espacial negativa (esquerda) e positiva (direita) usando o definição de adjacência de *rook*. O esquema de adjacência de *rook* considera unidades vizinhas aquelas que compartilham pelo menos uma aresta.



Moran's I global = -1 Moran's I global = 1

Fonte: Elaborado pelo autor (2023).

A partir dos *tweets* do município de São Paulo, o objetivo é verificar se os distritos com muitos ou poucos *tweets* estão próximos entre si. Duas variáveis quantitativas foram estabelecidas para a análise: i) o número absoluto de *tweets* e ii) o número de *tweets* em relação ao número de habitantes de cada distrito – *tweets* por habitantes, daqui em diante chamada de densidade de *tweets*. Para calcular a densidade de *tweets* o número de habitantes de cada distrito de São Paulo foi obtido do censo 2010 (Prefeitura de São Paulo, 2023).

Os índices de Moran global e local foram calculados para ambas variáveis. O índice local corresponde à autocorrelação em relação aos distritos vizinhos de primeira ordem (que compartilham fronteiras administrativas). Com o índice local é possível agrupar os distritos em cinco classes: LL (*Low-Low*), HH (*High-High*), LH (*Low-High*), HL (*High-Low*) e ns (*not significant*). O índice global corresponde à autocorrelação de maneira geral no espaço analisado e pode não ser representativo para capturar comportamentos localizados de autocorrelação.

Para determinar quais distritos tem uma estatística local significativa, um valor-p de 5% foi usado. Todo distrito com valor de autocorrelação espacial local com uma chance de ser obtido de maneira aleatória maior ou igual a 5% foi classificado como não significativo. A determinação dessa chance é feita por meio de diversas permutações dos valores para a estatística entre os distritos. Neste estudo, 999 permutações foram realizadas.

Para a determinação dos vizinhos de cada distrito, foi adotado o esquema de adjacência *queen*. O esquema de adjacência *queen* considera unidades vizinhas aquelas que compartilham pelo menos um vértice. As adjacências e pesos são definidas em uma matriz de pesos. Cada linha e coluna da matriz representa um distrito. Se um distrito é vizinho de outro, a respectiva linha e coluna recebe o peso 1. Caso contrário, recebe peso 0. A padronização por linha – em inglês, *row standardization* – foi aplicada a esses pesos. Nessa padronização, cada peso em uma linha é dividido pela soma dos pesos dessa linha. Assim, a influência do número de vizinhos de cada distrito não afeta de forma substancial o resultado final. Ou seja, há uma influência equilibrada entre distritos com mais e menos vizinhos.

VISUALIZAÇÃO DA DISTRIBUIÇÃO E AUTOCORRELAÇÃO ESPACIAL

Um arquivo *shapefile*, que contém informações geométricas, representou o mapa de São Paulo. Foi necessário cuidado ao renderizar visualmente o mapa e os *tweets* do município, visto que o sistema de coordenadas geográficas de ambas as informações deve ser o mesmo para não haver distorções.

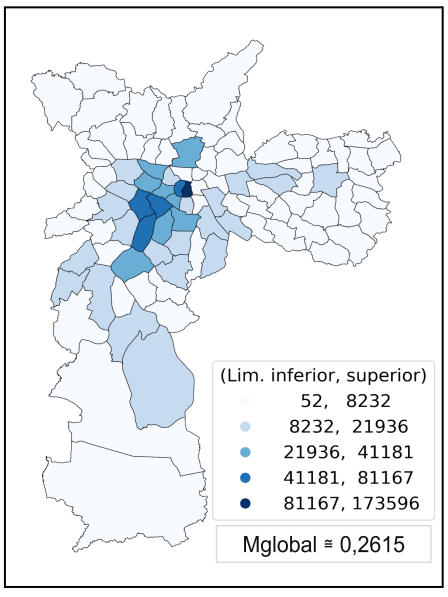
Para a visualização da distribuição espacial dos dados, mapas coropléticos foram utilizados. Nesses mapas, regiões são agrupadas em classes, e cada classe recebe uma cor. Por conveniência, o número de classes utilizado nos mapas plotados foi de cinco. O algoritmo usado para a determinação dos distritos de cada classe foi o de Fisher-Jenks; que retorna a melhor distribuição de elementos nas classes (Rey; Arribas-Bel; Wolf, 2020).

Os mapas de agrupamentos locais, para os quais a autocorrelação espacial local de cada distrito foi calculada, têm cinco classes, ao todo, por padrão.

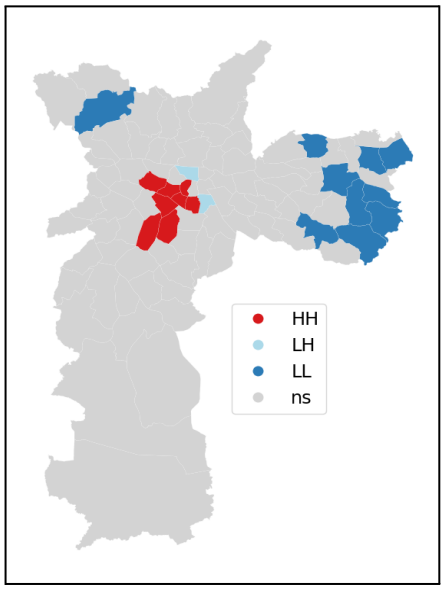
RESULTADOS

A Figura 3 mostra a distribuição espacial dos números para as variáveis número de *tweets* absolutos e densidade de *tweets*. A autocorrelação forte para o número absoluto de *tweets* significa que os distritos com muitos *tweets* ou poucos tweets estão próximos entre si.

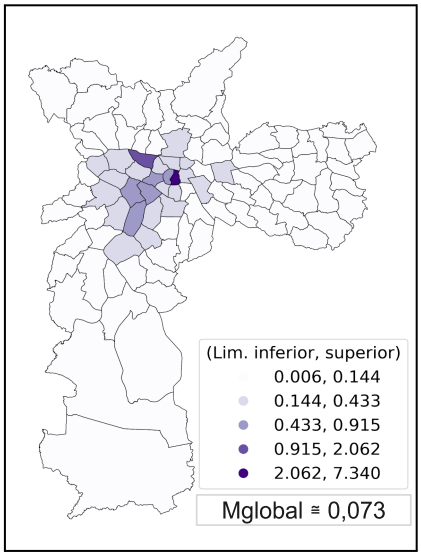
Figura 3: Distribuição espacial dos números para as variáveis número absoluto de *tweets* e densidade de *tweets*. a) distribuição do número absoluto de *tweets*. b) agrupamentos da autocorrelação espacial local do número absoluto de *tweets*. c) distribuição da densidade de *tweets*. d) agrupamentos da autocorrelação espacial local da densidade de *tweets*.



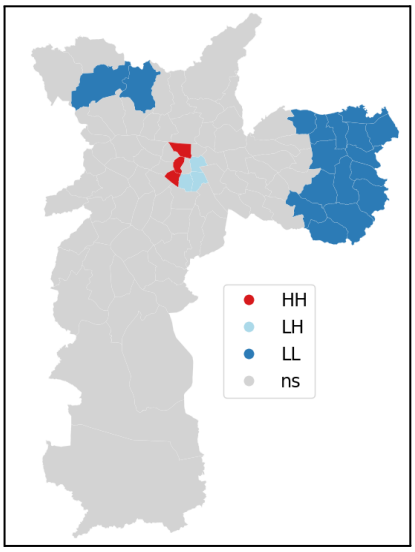
a)



b)



c)



d)

HH (*High-High*): distritos com autocorrelação alta e média dos vizinhos alta. LL (*Low-Low*): distritos com autocorrelação local baixa e média dos vizinhos baixa. LH (*Low-High*) distritos com autocorrelação baixa e média dos vizinhos alta. HL (*High-Low*): distritos com autocorrelação alta e média dos vizinhos baixa. NS (*not significant*): distritos com autocorrelação espacial não significativa.

Fonte: elaborado pelo autor (2023).

O índice de Moran local para ambas as variáveis variou entre positiva e negativa, dependendo do distrito (Figura 3-b e 3-d). Cada agrupamento no mapa revela que há um fenômeno em comum entre os distritos de cada grupo que faz esses distritos terem seus respectivos valores para a variável analisada.

CONCLUSÃO

Ao levar em conta o número absoluto de *tweets* presente em cada distrito do município de São Paulo, Brasil, a autocorrelação espacial global obtida pelo índice de Moran é positiva. Isso indica que distritos próximos possuem um número de *tweets* similar. Diferente do número absoluto de *tweets*, a densidade de *tweets* – *tweets* por habitantes – apresenta autocorrelação espacial global baixa. Porém, com a autocorrelação espacial local, foi possível localizar agrupamentos de distritos em que há uma autocorrelação espacial, tanto positiva, quanto negativa. A descoberta do fenômeno que causa a formação desses grupos é possível com posteriores análises, ao levar em conta outras variáveis de cada distrito, como informações demográficas, e de infraestrutura.

Agradecimentos

O autor agradece à Universidade Tecnológica Federal do Paraná pelo apoio financeiro, e ao seu orientador, Sidgley Camargo de Andrade, por oferecer uma excelente oportunidade de iniciação científica, e pela ajuda em momentos de dúvida.

Disponibilidade de código

O novo código do crawler está disponível em: https://github.com/fabio-bays/t-crawler.

Conflito de interesse

Não há conflito de interesse.

REFERÊNCIAS

ANDRADE, Sidgley Camargo de. The effect of intra-urban mobility flows on the spatial heterogeneity of social media activity: investigating the response to rainfall events. **International Journal of Geographical Information Science.** v. 36, p. 1140-1165, 3 ago. 2021. Disponível em: <https://doi.org/10.6084/m9.figshare.12921974>. Acesso em: 28 jun. 2023.

CHEN, Xiang; YANG, Xining. Does food environment influence food choices? A geographical analysis through “tweets”. **Applied Geography**, v. 51, p. 82-89, jul. 2014. Disponível em: <https://doi.org/10.1016/j.apgeog.2014.04.003>. Acesso em: 27 mar. 2023.

GUERRERO, Paulina et al. Revealing Cultural Ecosystem Services through Instagram Images: The Potential of Social Media Volunteered Geographic Information for Urban Green Infrastructure Planning and Governance. **Urban Planning**, v. 1, n. 2, p. 1-17, 6 jun. 2016. Disponível em: <https://doi.org/10.17645/up.v1i2.609>. Acesso em: 22 mai. 2023.

PREFEITURA DE SÃO PAULO. Dados demográficos dos distritos pertencentes às Subprefeituras. 2023. Disponível em: <https://www.prefeitura.sp.gov.br/cidade/secretarias/subprefeituras/subprefeituras/dados_demograficos/index.php?p=12758>. Acesso em: 04 ago. 2023.

REY, Sergio J.; ARRIBAS-BEL, Dani; WOLF, Levi J. **Geographic Data Science with Python**. 2020. Disponível em: <https://geographicdata.science/book/intro.html>. Acesso em: 01 jul. 2023.

WOOD, Spencer A. et al. Using social media to quantify nature-based tourism and recreation. **Scientific Reports**, v. 3, n. 1, 17 out. 2013. Disponível em: <https://doi.org/10.1038/srep02976>. Acesso em: 10 mai. 2023.

YANG, Wei; MU, Lan; SHEN, Ye. Effect of climate and seasonality on depressed mood among twitter users. **Applied Geography**, v. 63, p. 184-191, set. 2015. Disponível em: <https://doi.org/10.1016/j.apgeog.2015.06.017>. Acesso em: 5 mai. 2023.